#WHYIDIDNTREPORT: USING SOCIAL MEDIA AS A TOOL TO UNDERSTAND WHY SEXUAL ASSAULT VICTIMS DO NOT REPORT

by Abby Garrett

A thesis submitted to the faculty of The University of Mississippi in partial fulfillment of the requirements of the Sally McDonnell Barksdale Honors College.

Oxford May 2019

Approved by

_____
Advisor: Dr. Naeemul Hassan

_____
Reader: Dr. Carrie Smith

_____
Reader: Dr. Dawn Wilkins

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Naeemul Hassan for all his help throughout the research and writing of my Thesis. Dr. Hassan and the rest of the Computer Science Faculty have taught me so much during my time at the University, and I cannot express my gratitude enough.

Abstract

ABBY GARRETT : #WhyIDidntReport: Using Social Media as a Tool to Understand Why Sexual Assault Victims Do Not Report

(Under the direction of Naeemul Hassan)

Sexual assault has gone largely under-reported, and social media movements, like #WhyIDidntReport, have brought great awareness to this issue. In order to take advantage of the large amounts of data the #WhyIDidntReport movement has generated, the study uses tweets to explore reasons why victims do not report their assault. The thesis cites current research on the topic of assault to generate a list of explanations victims use to describe their lack of reporting and compares the distributions with existing studies. We use supervised learning technique to automatically categorize tweets into one of eight categories. This approach uses social sensing to determine why people do not report rather than surveys and interviews like current research.

The machine learning algorithms used to categorize the tweets as having a reason or not are Naive Bayes, Random Forest, and Recurrent Neural Networks. Only Naive Bayes and Random Forest were used for categorizing the reasons because there was not enough data to train large numbers of parameters of RNN. Each algorithm produces relatively precise results for the binary classification and categorizing whether a tweet references shame, denial/minimization, fear of consequences, hopelessness/helplessness, drugs or drinking or disassociation, lack of information, protecting the assailant, or age as the reason they did not report. These algorithms and tweets can be used to label data in future studies.

Using the current research, natural language processing, and machine learning, we were able to determine a list of reasons mentioned on Twitter under the #WhyIDidntReport movement. The distribution of the reasons differed from current research, most likely as a result of the form of data collection. However, the categories themselves were consistent with findings from other studies. The use of social sensing to determine reasons presents a new perspective on the topic and allows for comparison with other research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
CSV	Comma Separated Values
PTSD	Post Traumatic Stress Disorder
DDD	Drugged, Drunk, Disassociated

1 Introduction

1.1 Research Question

My thesis looks to determine if social media can be used to quantify the reasons victims do not report sexual assault. In order to answer this question, we look at why victims say they did not report in their tweets. Analyzing and summarizing the data allows it to be compared to existing literature to determine if the findings are accurate as compared to other conclusions drawn by answering the same questions in different ways.

1.2 Background

The idea for the topic originated from the social prevalence of the issue of sexual assault. The #MeToo movement has been a powerful tool to show people just how large of an issue assault is. The movement has opened a national dialogue about how to address this issue and help victims realize that they are not alone.

The #MeToo movement was the first to call social media users to action to express how common sexual assault is in the world. When looking through the tweets and reading articles discussing the movement, I became increasingly interested in why victims did not report these incidents. Once this more narrowed topic was selected, we discovered the #WhyIDidntReport movement which was originally started to support Dr. Ford. She is the witness who claimed Supreme Court Judge Kavanaugh attempted to assault her with a friend at a party when they were in high school. After Trump tweeted, "I have no doubt that, if the attack on Dr. Ford was as bad as she says, charges would have been immediately filed with local Law Enforcement Authorities by either her or her loving parents. I ask that she bring those filings forward so that we can learn date, time, and place!" [Noveck2018]. Other victims came forward citing their own instance of not reporting using #WhyIDidntReport. They came forward to contradict the argument that Dr. Ford was lying because she never said anything about it prior to Kavanaugh's nomination thus making it all a

political ploy. Women from around the world started flooding Twitter with their own stories of why they did not report in order to suggest that Dr. Ford is not alone in keeping her experience to herself. My thesis does not look to address the validity of her claim, but rather to utilize the vast data her accusations spurred to see if a conclusion can be reached on why victims do not report the assault and to see if these conclusions align with other research on the issue.

The goal of this thesis is to have a better idea of why victims of sexual assault do not report their experiences. Understanding why victims do not report is an important part of solving the problem so action can be taken to change the culture surrounding reporting and sexual assault. Novek references statistics gathered from the Justice Department stating that 7 in 10 victims do not report their assault [Noveck2018]. Because of the lack of reporting, the numbers of how often sexual assault occurs may have been skewed. Fortunately, social media movements, like the #Metoo movement, have highlighted what statistics did not, resulting in improved societal awareness.

Current research seeks to detail the reasons victims do not report through surveys and empirical research. Many surveys are limited in size and focus on specific subsets of victims. In "Why Sexual Assault Survivors Do Not Report to Universities: A Feminist Analysis" the study draws on 220 responses and focuses on college women [Spencer et al.2017]. National surveys like [Fisher et al.2003] and [Langton and Truman2015] are able to collect larger samples of data, but both still ask pointed questions that could result in answers the subject may not have come to on their own, which differs from self-reported data. They also focus on a specific sample of the population. Empirical research offers a different approach, but as seen in [Engel2017], it can result in a more informal study. Each methodology presents certain restrictions and benefits, just as this study does. Using Twitter and social movements allows for a larger sample, but it cannot be as detailed and allows the victim to be affected by the response of others.

1.3 Contributions

My thesis draws on other research for the sake of comparison and generating a list of reasons victims do not report. I compare my findings to other work in order to see how collecting data in different ways affects results. Other research is used to establish reasons victims do not report because the main focus of this thesis is to see how reliable Twitter data is on the topic of reporting sexual assault. It establishes a different way of studying the lack of reporting because Twitter uses self-reporting, is a public platform, and presents a vast and varied sample.

Machine learning algorithms were used to best present the findings collected from Twitter and test the accuracy of machine learning algorithms on this kind of data. The algorithms were generated to categorize tweets as mentioning or not mentioning a reason for not reporting the assault which is known as binary classification because it is concerned with two classes. It then goes one step farther to label the tweets with which reason was mentioned or implied which is multi-class classification since it classifies based on many classes.

The research revealed a difference between survey reported data and self-reported data. Survey reports showed a higher appearance of denial and minimization, while my research, using self-reporting, more often cited a belief that nothing would or could be done or a fear of the consequences of reporting.

For the technical aspect of this research, machine learning was successfully implemented to categorize tweets concerning sexual assault reporting. Naive Bayes, Random Forest, and Recurrent Neural Networks were all relatively precise at predicting whether or not a tweet mentioned a reason. Each machine learning algorithm uses different techniques to classify data. Similar to different ways of collecting data, the algorithms each present restrictions and benefits. However, only Naive Bayes and Random Forest were successful in labeling which reason was generated because Recurrent Neural Networks require more data when more variables are being considered.

All aspects of the study combined to show that Twitter presents an interesting avenue for discovering why victims do not report. Its unique platform for victims to be able to self-report and the numerous users and data points allow for a new way of answering a highly relevant social question. It presented new findings on what reasons are most common to victims rather than exploring new reasons for not reporting.

2 Related Works

2.1 Sexual Assault

Sexual assault is a huge problem that greatly affects the victims throughout their lives. [Ullman2016] studied whether childhood assault was related to PTSD or problem drinking later in life. The study utilized a mail survey and concluded that childhood assault can result in PTSD about 30% of the time and problem drinking 20%. Sexual assault is not an isolated event that victims experience and they move on from. It can dictate how they perceive and interact with the world around them for the rest of their lives. [Ullman and Peter-Hagene2014] discovered that when victims receive supportive reactions when they disclose their experience, they feel more control over their own recovery. The mail survey indicated that most victims told at least one person and having support resulted in the use of more positive coping mechanisms.

The studies were done on sexual assault reveal a negative impact on victims lives. Victims have methods available to deal with the trauma. Hot lines and various forms of therapy can help a victim learn to cope with what happened to them. However, having such a traumatizing experience affects many victims for the rest of their lives. The best way for society to protect victims is to do everything to prevent future victims. Understanding how prevalent the issue is and making sure victims feel as safe as possible disclosing their experiences are both positive first steps in creating a better culture surrounding assault victims.

2.2 Reasons for Not Reporting

Related research that has been conducted on why victims do not report assault takes a more narrowed approach. The approaches are one way of answering the question of why victims do not report and each have their own sets of advantages and limitations just as this study does. Many look at specific subgroups of victims or reporting, and they predominately use surveys. Two studies both looked into why college women did not report their assault by surveying college women nationally

or by surveying women on a specific college campus [Fisher et al.2003], [Spencer et al.2017]. The National Survey focused on college women reporting to people other than authorities by using a survey with 4446 responses which was the largest sample size discovered from a survey research project [Fisher et al.2003]. The University focused survey was interested in why women do not report to universities specifically and had a sample size of 220 [Spencer et al.2017]. Both studies did reach similar results as to why, in these cases women, rarely report. They reference that the women were afraid, ashamed, did not think it was bad enough, or were drugged or drinking ([Fisher et al.2003], [Spencer et al.2017]). The national survey also mentions a relationship to the offender and personal attributes like race, age, and gender can make a woman less likely to report [Fisher et al.2003]. The college women survey brings up women did not know they could report, did not want to get the offender in trouble, or did not report to the university because it was not university-related [Spencer et al.2017].

[Mengeling et al.2014] focus specifically on service women and also uses a survey which was mailed to participants. The study drew on about 2800 participants which included those that chose not to respond. It presents another example of a larger survey that again narrows the focus to a specific group of people. However, [Khan et al.2018] present a slightly different approach. The study is slightly different than the topic of this thesis as it focuses on why victims may not label their experience as assault. However, in the end, the result focuses on how this can affect reporting. It brings a new approach because [Khan et al.2018] uses a mixed form of data retrieval. The study gathers data using interviews, focus groups, and observation of subjects. It also looks at undergraduates, but the difference in data gathering presents a new approach to the subject.

Yet another approach, [Ménard2005] studies the question using data gathered from 48 rape crisis centers all in Pennsylvania. This system differs from the others in that it is more of a voluntary form of sharing, but it is still limited in location. Ménard does have another aspect of the study that focuses more on cultural effects of not reporting, but because the second part of the study focuses on the individual which is what my thesis focuses on, this is the part that will be used throughout this report [Ménard2005].

Dr. Beverly Engel wrote an article published in Psychology Today that uses empirical data to compose a list of reasons victims do not report. Dr. Engel does not reference a specific study, but rather, pulls on her experiences as a psychologist helping victims deal with their abuse to present reasons she has found that explain why victims do not report [Engel2017]. She is widely known as

being an advocate for sexual assault victims and has written twenty-two books to aid victims. This credibility gives her a large amount of experience to draw on in establishing seven reasons she has found why victims do not report. Engel explains eight reasons common for not reporting among her patients.

1. Shame
2. Denial or minimization
3. Fear of consequences
4. Low self-esteem
5. Feelings of hopelessness and helplessness
6. History of being sexually violated
7. Lack of information
8. Disbelief, drugged, or disassociated

Her article was used heavily in deciding on which reasons to focus on when looking into the Tweets because she does a good job outlining exactly what falls into each category of reasons. Her article became kind of the framework to work off of when composing the final list of reasons and was backed up with conclusions from other studies [Engel2017].

Using the research discussed thus far, we composed a list of eight reasons with which we categorized the tweets. Each reason appeared not only in several papers with relevant current research but also was found to occur often when manually going through the tweets. Each tweet can fall into multiple categories. It simply depends on how many reasons the victims references in their text. The categories used are detailed below.

1. Shame: the victim feels it is their fault, they were too ashamed to tell anyone, or they felt damaged because of what happened
2. Denial/Minimization: trying to convince themselves that what happened really was not a big deal, did not happen, or was not wrong, or trying to forget what happened and feeling as though reporting makes it real

3. Fear of Consequences: This category is rather broad because it can be a fear of losing one's job, what people will think, the actual legal aspects of reporting, physical harm from the assailant or other, or anything else that could cause the fear which prevents the victim from reporting. The main point to make here is that fear of not being believed does not fall into this category because it is detailed in another category and we do not want to see overlap between categories for the sake of the machine learning algorithm.

4. Hopelessness/Helplessness: Victims feel as though there is no point reporting because they have seen how people treat those who do report and have seen the lack of action when an assault is reported, especially due to others' disbelief.

5. Drugged/Disbelief/Disassociation: substance or psychological effects prevent the victim from having a clear memory of the events

6. Lack of Information: not realizing where, how, or who to report to or simply not realizing that they can report the assault

7. Protecting Assailant: some victims do not want to see their assailant go to jail or have their life ruined, sometimes through the persuasion of others

8. Age: when victims cite that they did not report because they are young

2.3 Related Social Movements

Social media presents a new way for many people to feel as though they have a voice. Social movements have mobilized this ability with hashtags that create a resounding voice across platforms.

The Black Lives Matter movement was one that generated a voice on social media. Carney looks at tweets that flooded Twitter following the deaths of Michael Brown and Eric Garner in 2014 and the related trials [Carney2016]. It discusses the fact that social media can help shape views, but also allows for interactions and engagement with content. The results reveal that Twitter was a good way of assessing the societal reaction to the trails mentioned in the paper and the movement as a whole, but also the length restriction forces users to choose their words carefully [Carney2016]. While the movement can be affected and framed by social media, [Ince et al.2017] discusses that many of the tweets surrounding #BlackLivesMatter were positive or expressed solidarity. Part of the reason could be connected to the social perception of disagreeing. Ine, Rajos, and Davis also express how Twitter blows up with these movements when other events spark awareness in [Ince et al.2017].

Black Lives Matter brought attention to racism within the police force. The Me Too movement sparked universal recognition of the prevalence of sexual assault. Victims everywhere started expressing their own experiences or at least citing that they too had been assaulted. #MeToo offers the opportunity to bring awareness and change to the issue. In the journal *The Lancet*, [O'Neil et al.2018] express the hope that the movement will push for sexual harassment to be considered a health issue thus offering health-related support. Studies like [Ward2018] reference that simply making people aware of the issue has positive impacts to remind people to be cognizant of how their actions can make others feel. The social media movement really highlighted how widespread the issue of assault is.

Social media and especially social media movements bring awareness to societal situations. The #WhyIDidntReport movement focuses specifically on how common it is for victims not to report their assault. As seen with studies of Black Lives Matter and Me Too, social media can be a good representation of societal perception surrounding given topics, as well as help shape what said perception is. Social media is a powerful tool presenting vast and varied data on hot topics of conversation taking over societal topics.

3 Method

The research was conducted using data from Twitter. Taking this avenue allowed for a data set that included geographic and user variety without sacrificing a large amount of time. Web scraping gathered the data which was then manually labeled in order to allow for later labeling the data with machine learning algorithms. Naive Bayes, Random Forest, and Recurrent Neural Networks were each used to classify data due to their varied techniques and resulted in reliable classification.

3.1 Data Retrieval

Twitter data has been utilized previously in other studies in order to analyze the reasons why victims of sexual assault decide not to report. In order to accumulate enough data, we ran a web scraping program written by Dr. Naeemul Hassan to compile files of tweets that included #WhyIDidntReport. The program pulled approximately 40,000 tweets which contained a variety of subject matters that ranged from commentary on the Kavanaugh hearings to personal accounts of assault to reasons why victims did not report the assault.

3.2 Data Description

Because the data was pulled from Twitter, it has a variety of users and places. Table 1 details that the data collected represents approximately 25,000 unique users, 211 cities, and 43 states. Because Twitter is a public platform, it can draw on a vast number of people from all over the nation. It allows for so much more data to be taken into consideration than a survey, interview, or empirical data retrieval in a reasonable amount of time. The other techniques can allow for just as large of a sample size, but social media allows for more data to be gathered quickly.

The sample was gathered over the course of 11 days, from September 22, 2018, to October 2, 2018. The range aligns perfectly with the beginning of the #WhyIDidntReport movement which began as a result of President Trump's tweet on September 21, 2018. This tweet sparked a massive

Table 1: Data Statistics

Total Days Collected	11
Unique Users	25,000
States Represented	43
Cities Represented	211

influx of data surrounding the reasons victims do not report the assault, so collecting data while that movement was fresh was the best opportunity for a large sample size. Table 2 also shows the range of lengths of tweets. An important note to make about the data is that tweets can be no longer than 280 characters, thus limiting the amount of information we are able to glean from any given account.

Table 2: Data Ranges

	Min	Max
Days Collected	9/22/2018	10/2/2018
Tweet Length (characters)	18	280

3.3 Data Labeling

After labeling about 1000 tweets, it was discovered that the list of tweets included too many that did not actually share reasons the victims chose not to file a report. Because we originally wanted more tweets that included a reason, I wrote a python program that narrowed down the tweets to approximately 10,000. The python program searched each tweet for keywords or phrases that generally indicated a tweet included a reason the victim did not report. In order to decide which phrases to search for, I manually labeled 1,000 tweets with whether or not the victim mentioned a reason they chose not to report. If the tweet was labeled with containing a reason, I marked the words or phrases that usually triggered the statement of explaining why. The phrases chosen were biased to my perspective, but the purpose was simply to narrow down the data to more tweets without reasons and the phrases chosen served this purpose well. After labeling, it was discovered that most of the tweets that were useful for the research being conducted contained "because", "I was", "I thought", "I felt", or "I didn't". More phrases could have been chosen, but these tended to indicate a reason the most often. If the tweet did contain one of these phrases, then the program added the information to a data frame which was then written to a CSV file.

The filtered tweets were used to train machine learning algorithms. In order to have training and testing data, I manually labeled approximately 8,000 tweets with whether or not the tweet did, in fact, contain a reason. Figure 1 shows that approximately 85% of the tweets contained reasons. When training the machine learning algorithms, more tweets without reasons were needed in order to accurately train on both categories, so I manually labeled more tweets in order to have a balanced distribution. The new distribution contained 4000 tweets with reasons and 4000 without reasons, thus supplying data with a fifty-fifty distribution.

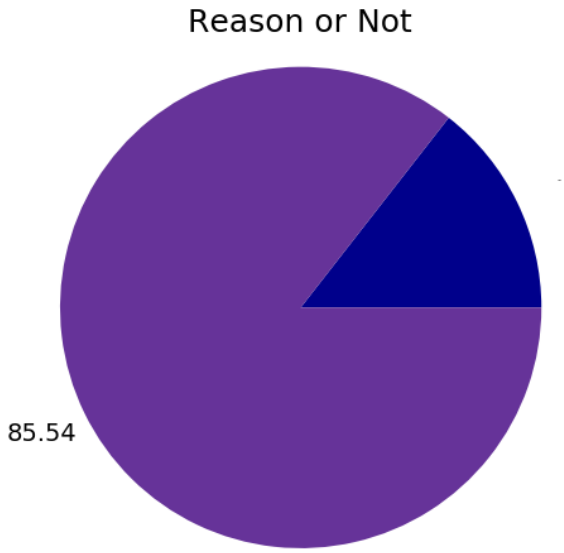

Figure 1: The graph portrays the distribution of whether or not the original tweets contained a reason.

If the tweets did include reasons, then I labeled it with which reasons the victim mentioned. The reasons that were considered for labeling were accumulated using psychology research and manually seeing which were most relevant in the tweets. Figure 2 reveals the distribution of the reasons across the tweets. Each tweet could be labeled with more than one reason, but as seen in Figure 3, the majority only contained one or two reasons. This also only showcases the reasons defined for this study, but other reasons could have been mentioned in the tweets. In Figure 3, the column representing 0 reasons means that a reason was mentioned, it just did not fall into one of the categories defined. From the graph, we see that this did not occur often.

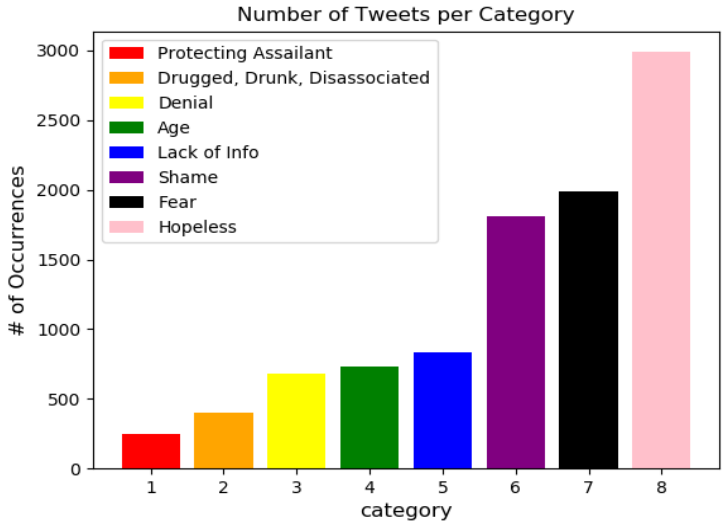

Figure 2: The graph portrays the number of tweets that mentioned each reason.

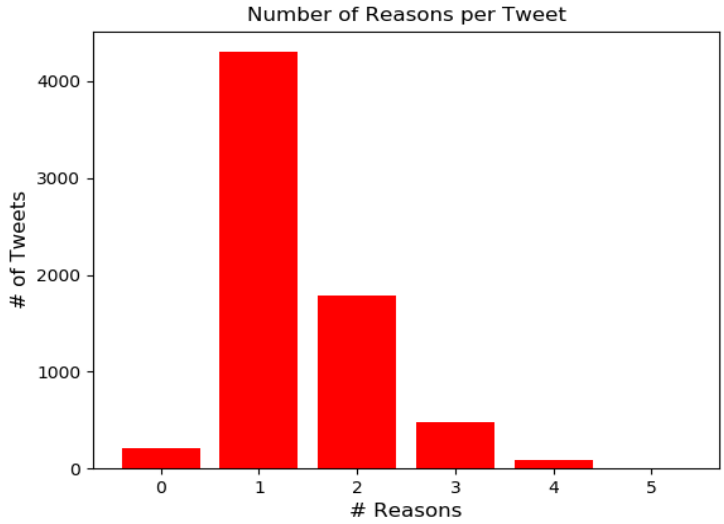

Figure 3: The graph portrays how many reasons were mentioned in each tweet.

3.4 Classification

Once the tweets had all been labeled the actual classifying of the data began. The data was labeled with two things, whether a reason was mentioned or not and the reasons that were mentioned. We ran binary classification tests, labeling having a reason or not, with three machine learning algorithms: Naive Bayes, Random Forest, and Recurrent Neural Networks, and then ran conducted multi-class classification, the specific reasons mentioned using only Naive Bayes and Random Forest.

The multi-class labeling labeled tweets with only one reasons, even though it may have contained multiples.

3.4.1 Overview

We chose to use three classification models in order to see which was the most precise for labeling the set of data for both binary and multi-label classification. The models all classify data using different techniques, so in order to present the best analysis, it was necessary to compare a variety of models.

For each of the classification models, we tested performance based on the number of tweets that were fed into the classification model. The testing and training sets were taken by stratifying the data to represent the uneven spread of categories. The research used count vectors, word TF-IDF's, and n-grams TF-IDF's as the features. Features are what the algorithm considers to determine the probability that a tweet falls into a given class.

1. Count: the count of the number of times a word is used throughout all tweets

2. Word TF-IDF: The basic idea is that the word TF-IDF takes into account how often a word is used and gives more weight to words used less as they can more often signify which class a tweet falls into.

3. N-gram TF-IDF: This is similar to Word TF-IDF, but rather than use words, it uses a string of words. In our case, we considered 1-3 words in each string.

For word and n-grams, we can determine the number of features we want to consider. Cutting this off at a specific number limits the number of most frequently used words to consider. We tested a variety of numbers of features but did not see any significant variation so ended up using 7000 features, meaning we look at the 7000 most commonly used words or n-grams in the tweets.

3.4.2 Naive Bayes

The Naive Bayes classification model uses features and Bayes formula to determine which category a given tweet falls into. It is said to be naive because it assumes independence of the features in its determination. The assumption of independence poses problems because the assumption is rarely completely true. However, Naive Bayes was chosen because it does not need much data to be trained well, is easy to implement, and is a popular algorithm for text classification. How little data

is needed to train Naive Bayes also does not prevent it from being accurate with larger amounts of data points [Soni2018].

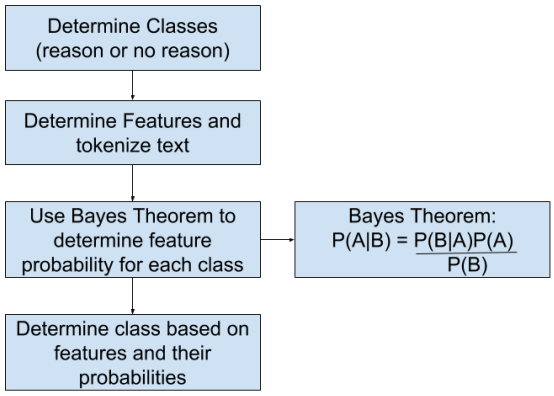

Figure 4: The figure summarizes the steps of the Naive Bayes classification.

In this study, we use the word count vector, the word level TF-IDF, and n-gram TF-IDF as the features. We use the 7000 most common words and n-grams in determining which class a tweet falls into.

3.4.3 Random Forest

Random Forest is an algorithm that generates a forest of decision trees. Forest really just means many decision trees that it then averages together. Decision trees work similarly to the way they sound. The algorithm takes into account the outcomes of certain events to decide what step to take next. The decision trees are an important part of the Random Forest Model, so Figure 5 is a basic example of a decision tree.

Random Forest looks at the weight of importance of each feature in deciding which ones to drop upon training and allows for a lower variance by considering a forest of decision trees. My model averages 10 decision trees which is the default for the algorithm. We could have considered using a

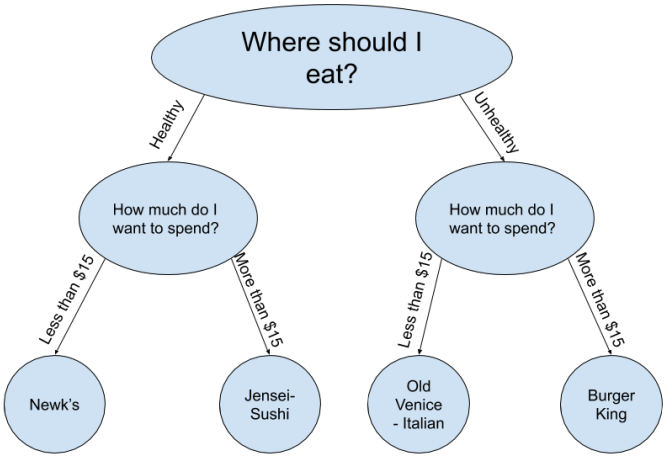

Figure 5: The figure details a decision tree example.

different number, but since the results shown with the default were good, there did not seem to be a need.

3.4.4 Recurrent Neural Networks

Recurrent Neural Networks work kind of like a brain because it learns from its mistakes. The training set is run through the algorithm then tested against the labeled data. The data is sent back through the algorithm to learn from its mistakes. The way this algorithm label is using an embedding layer. Embedding layers allow for the algorithm to look at more than just the word count. Embedding layers allow for a more detailed look at the text in order to better understand what makes each tweet fall into each class. Figure 6 depicts the process of the data being run through the algorithm, and how the data is pumped in a cycle in order to allow for the recurrent aspect mentioned earlier.

For my algorithm, the embedding vector is of length 32. This means that the algorithm is able to break the text into a vector containing 32 statistics which allows for the more detailed look at

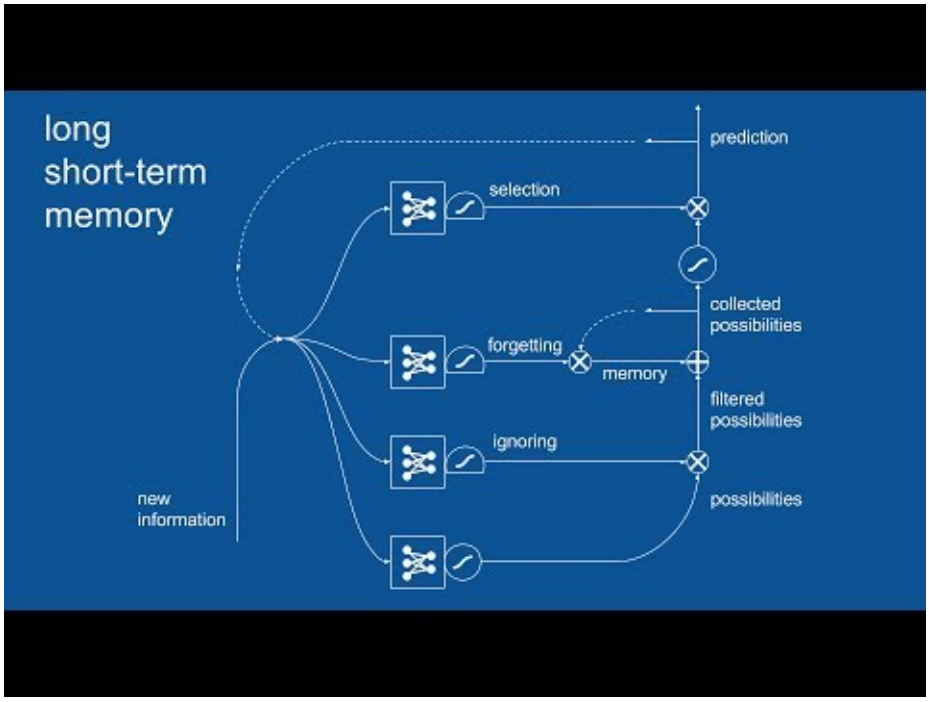

Figure 6: The figure details the RNN LTSM model [Rohrer2017]

why each tweet falls into a given class. I also have 100 embedding layers and 3 epochs which were determined to be the most efficient due to the time taken and accuracy. I also used a Long-Short Term Memory RNN which is just a specific type of RNN. The LTSM RNN model allows for the RNN to remember more than just the most recent step, therefore better being able to predict patterns.

My RNN was very accurate in binary classification, but it takes large amounts of data to train properly. Therefore, the RNN was only used for binary classification rather than both binary and multi-class since not all classes had significantly large training or testing data.

4 Results

4.1 Classification Algorithms

4.1.1 Binary Classification

1. Unbalanced

   The binary classification at first produced highly precise results. To begin, the data was not balanced, so the precision with which the algorithm appeared to label the data was shown to be because the algorithm was labeling almost everything with having a reason. Table 3 shows the classification report associated with Naive Bayes using the count vector as the features. Figure 7 graphs the precision of both Naive Bayes and Random Forest using the count vector. The comparison of the graph and classification report revealed that the graph is not an accurate representation of how effective the algorithm was.

Table 3: Naive Bayes, Unbalanced, Feature: Count Vector

	precision	recall	f-score
no reason	0.67	0.48	0.56
reason	0.92	0.96	0.94
micro avg	0.89	0.89	0.89
macro avg	0.79	0.72	0.75
weighted avg	0.88	0.89	0.88

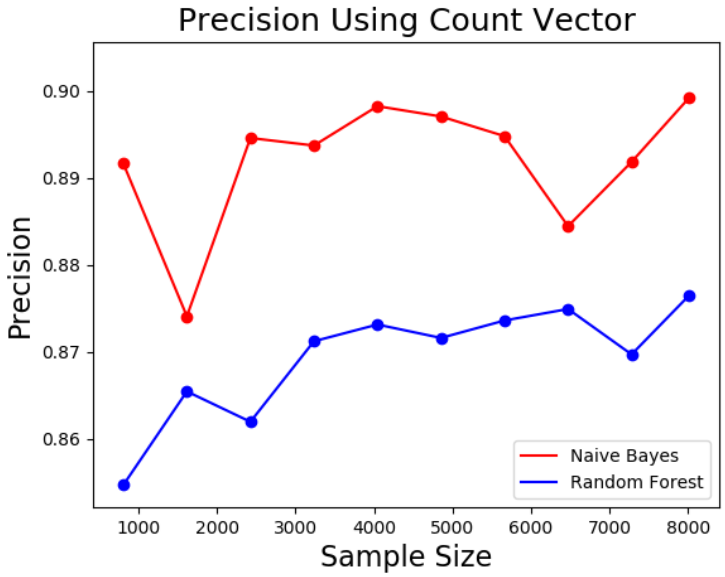

Figure 7: The graph portrays the precision of the classification algorithms using count vectors.

Figure 8 details the classification algorithms using TF-IDF word vectors. RNN performed the best with almost 89% precision, but Naive Bayes and Random Forest were not far behind in precision. Again, the classification report, detailed in Table 4, reveals the average precision was not a fair representation because the algorithm was unevenly labeling the data since the data itself was unbalanced.

Table 4: Naive Bayes, Unbalanced, Feature: Word TF-IDF

	precision	recall	f-score
no reason	0.00	0.00	0.00
reason	1.00	1.00	0.92
micro avg	0.86	0.86	0.86
macro avg	0.43	0.50	0.46
weighted avg	0.73	0.86	0.79

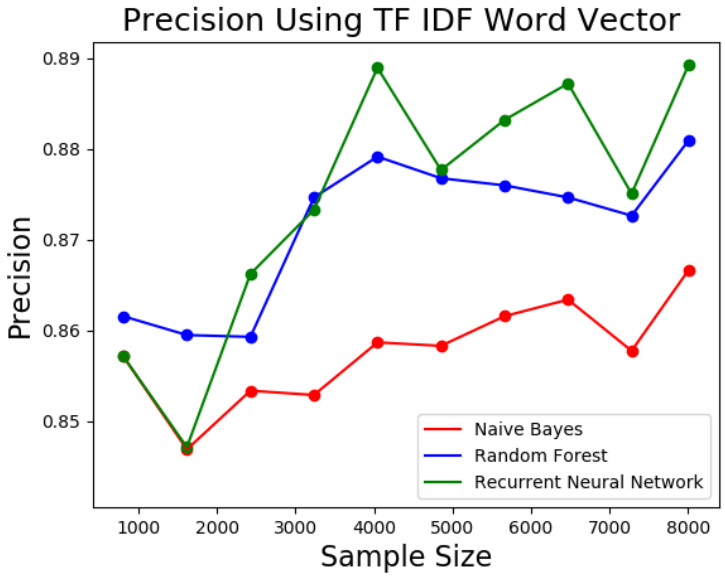

Figure 8: The graph portrays the precision of the classification algorithms using TF-IDF word vectors.

Figure 9 portrays the results of using 1-3 length word grams as the features. The slope stays positive fairly consistently as the number of tweets used increased. It also reveals no clear better classification algorithm, but it appears as though Naive Bayes was surpassing the precision of Random Forest more clearly as the number of tweets continued to increase. Yet again, the classification report shown in Table 5, details that the algorithms were not labeling both categories well, but rather labeling reason so well that it increased the average precision since most tweets contained a reason.

Table 5: Naive Bayes, Unbalanced, Feature: N-gram TF-IDF

	precision	recall	f-score
no reason	0.67	0.48	0.56
reason	0.92	0.96	0.94
micro avg	0.89	0.89	0.89
macro avg	0.79	0.72	0.75
weighted avg	0.88	0.89	0.88

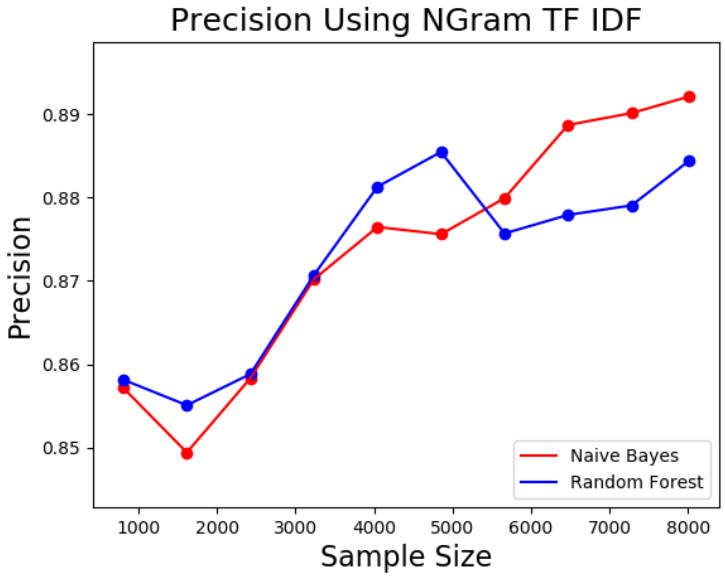

Figure 9: The graph portrays the precision of the classification algorithms using word n-grams.

While the results look very promising, after viewing the classification reports it was evident that much of this was due to the algorithms classifying almost everything as having a reason. To train the algorithms more precisely across the board, we went back and added more tweets without reasons, so the spread was more balanced.

2. Balanced

We adjusted the data to include 4000 tweets with reasons, and 4000 tweets without reasons. Once the data was balanced, the classification algorithms, started to produce more precise results across both classes. The same features and algorithms were used to conduct binary classification on the unbalanced and balanced data. The classification report for Naive Bayes is the only one listed because the classification reports did not vary greatly between algorithms. The main purpose is to show that the algorithms started to produce better results across both classes.

(a) Feature: Count Vector

While the results do not produce much better results overall once the data starts leveling out, the overall scores are better. Table 6 shows that using balanced data presented an algorithm that was efficient for both classes, not just one. Using the count vector as the feature, Naive Bayes performs better, leveling out at about 90% precision.

Table 6: Classification Report: Naive Bayes, Balanced

	precision	recall	f-score
no reason	0.95	0.85	0.90
reason	0.86	0.96	0.91
micro avg	0.90	0.90	0.90
macro avg	0.91	0.90	0.90
weighted avg	0.91	0.90	0.90

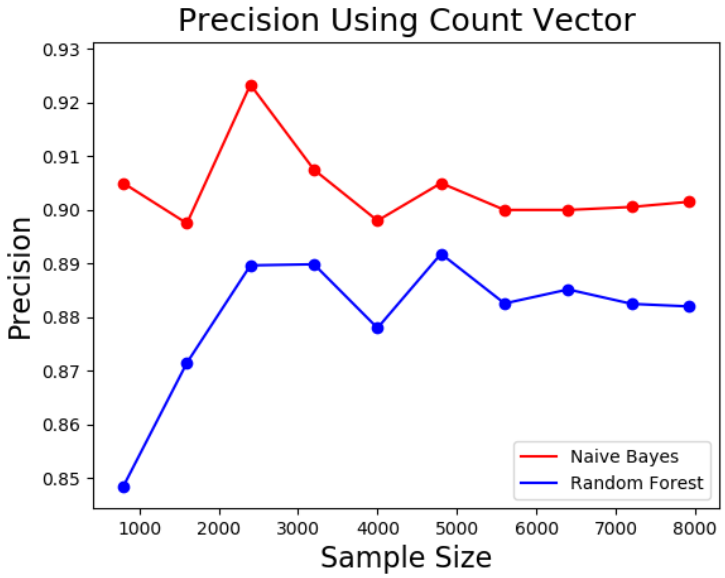

Figure 10: The graph portrays the precision of the classification algorithms using count vectors and balanced data.

(b) Feature: Word TF-IDF

The classification report shown in Table 11 again shows that using balanced data again provided better results for both classes. The overall results are slightly better, with all three algorithms providing about 90% precision using balanced data. RNN is also seen to slightly outperform the other algorithms.

Table 7: Classification Report: Naive Bayes, Balanced

	precision	recall	f-score
no reason	0.95	0.83	0.89
reason	0.85	0.96	0.90
micro avg	0.89	0.89	0.89
macro avg	0.90	0.89	0.89
weighted avg	0.90	0.89	0.89

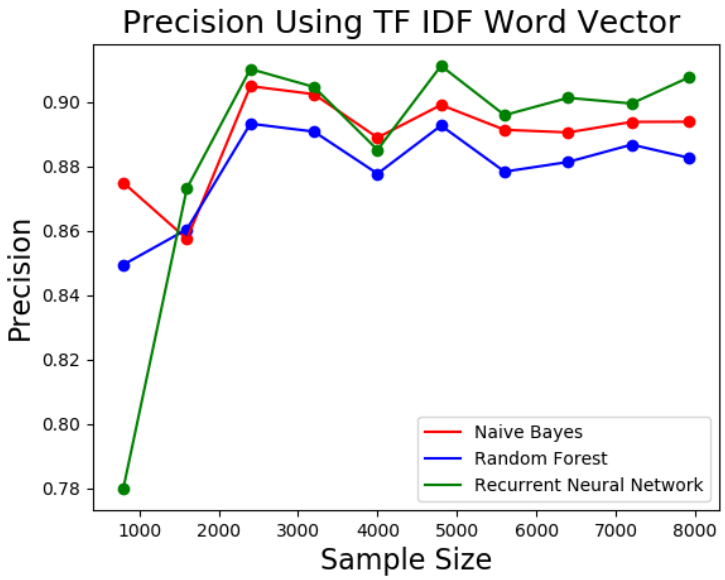

Figure 11: The graph portrays the precision of the classification algorithms using TF-IDF word vectors to classify balanced data.

(c) Feature: N-gram TF-IDF

Using balanced data slightly improves average precision when using n-grams as the feature, but again we see the classification report greatly improves. As a whole, n-grams perform roughly the same as the other features when applied to balanced data for binary classification.

Table 8: Classification Report: Naive Bayes, Balanced

	precision	recall	f-score
no reason	0.94	0.86	0.90
reason	0.87	0.94	0.91
micro avg	0.90	0.90	0.90
macro avg	0.90	0.90	0.90
weighted avg	0.90	0.90	0.90

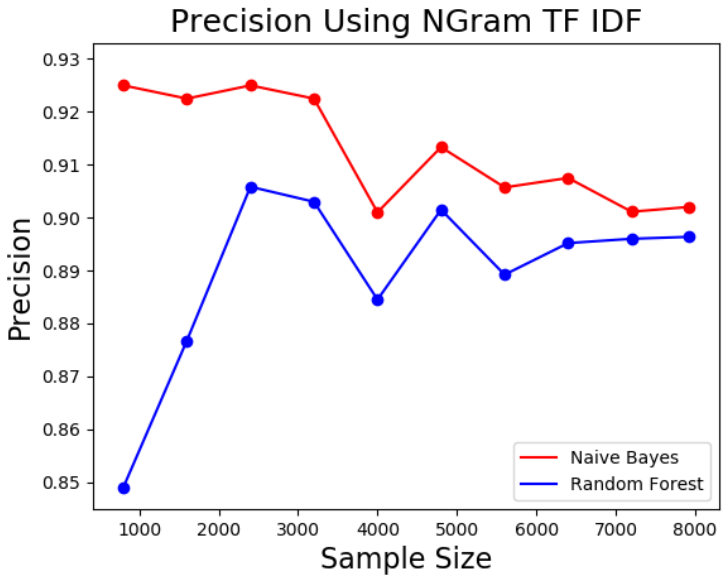

Figure 12: The graph portrays the precision of the classification algorithms using word n-grams to classify balanced data.

4.1.2 Multi-class Classification

The multi-class classification produced good results at first. Only Naive Bayes and Random Forest were used in this classification. Naive Bayes generally outperforms Random Forest as far as maximum precision. The testing and training data were taken using stratified random samples in order to account for the variety of appearances of each reason.

1. Unbalanced

   The below results detail how the algorithms performed when using the three features described earlier in the paper. The overall idea is that while the algorithms appeared to perform well, the classification algorithms reveal that the average was not a fair representation of the overall performance. Some classes were labeled with high precision, but others, like Protecting Assailant were never classified correctly, as seen in all three classification reports.

(a) Feature: Count Vector

Table 9: Naive Bayes, UnBalanced, Multi-Class

	precision	recall	f-score
Shame	0.38	0.47	0.42
Denial/Minimization	0.29	0.06	0.10
Fear of Consequences	0.38	0.42	0.40
Hopelessness/Helplessness	0.46	0.70	0.56
Drugged, Drunk, Disassociated	0.40	0.04	0.08
Lack of Information	0.39	0.20	0.27
Protecting Assailant	0.00	0.00	0.00
Age	0.25	0.05	0.08
micro avg	0.42	0.42	0.42
macro avg	0.32	0.24	0.24
weighted avg	0.38	0.42	0.37

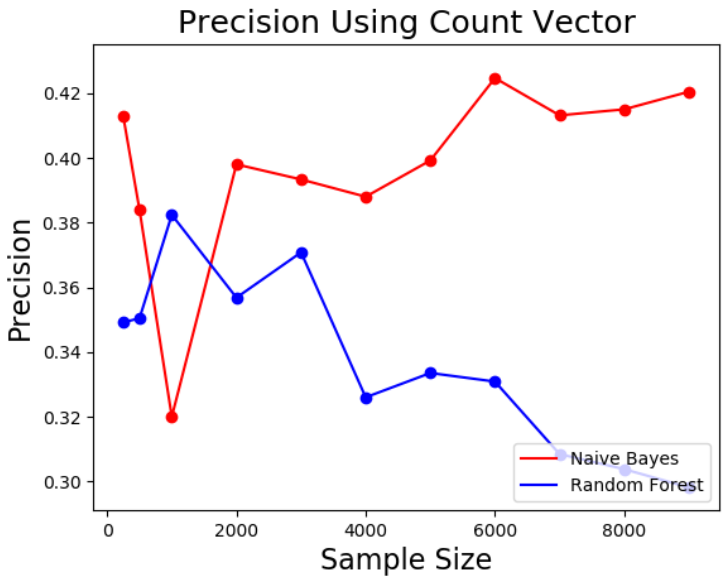

Figure 13: The graph portrays the precision of the multi-class classification algorithms using count vectors.

(b) Feature: Word TF-IDF

Table 10: Naive Bayes, UnBalanced, Multi-Class

	precision	recall	f-score
Shame	0.45	0.11	0.18
Denial/Minimization	0.00	0.00	0.00
Fear of Consequences	0.43	0.10	0.16
Hopelessness/Helplessness	0.33	0.98	0.50
Drugged, Drunk, Disassociated	0.00	0.00	0.00
Lack of Information	0.00	0.00	0.00
Protecting Assailant	0.00	0.00	0.00
Age	0.00	0.00	0.00
micro avg	0.34	0.34	0.34
macro avg	0.15	0.15	0.10
weighted avg	0.27	0.34	0.22

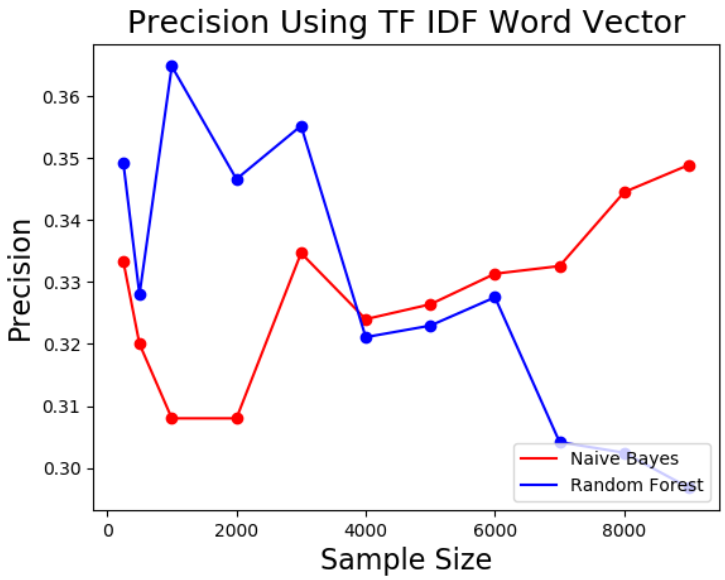

Figure 14: The graph portrays the precision of the multi-class classification algorithms using TF-IDF word vectors on the reasons mentioned the least.

(c) N-gram TF-IDF

Table 11: Naive Bayes, UnBalanced, Multi-Class

	precision	recall	f-score
Shame	0.46	0.43	0.44
Denial/Minimization	0.12	0.01	0.01
Fear of Consequences	0.48	0.43	0.45
Hopelessness/Helplessness	0.43	0.84	0.57
Drugged, Drunk, Disassociated	0.50	0.01	0.02
Lack of Information	0.39	0.13	0.20
Protecting Assailant	0.00	0.00	0.00
Age	0.50	0.05	0.08
micro avg	0.44	0.44	0.44
macro avg	0.36	0.24	0.22
weighted avg	0.42	0.44	0.37

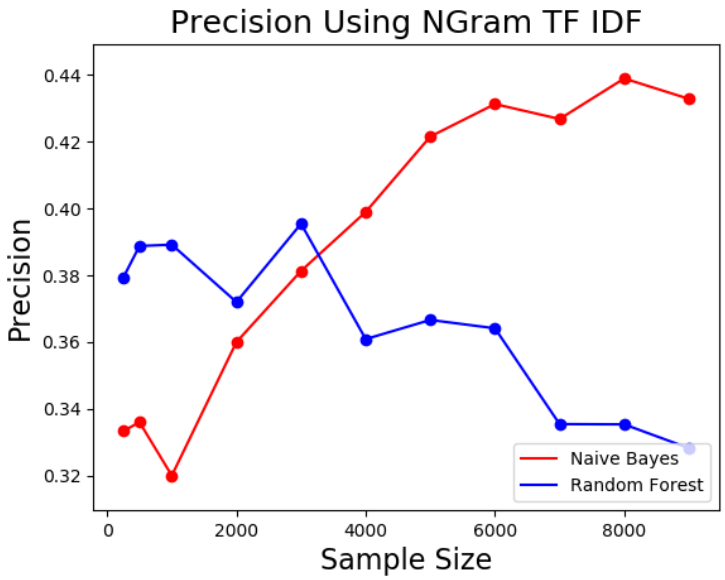

Figure 15: The graph portrays the precision of the multi-class classification algorithms using ngrams.

After examining the results of the algorithms, I generated a confusion matrix which exemplifies why the data needs to be balanced into groups with similar amounts of data. We want the number to be closest to one on the diagonal, but because not enough data is present, this is not always the case. False negatives are given by the column, and false positives by the row.

Table 12: Naive Bayes, Unbalanced, Multi-Class Confusion Matrix

	Shame	Denial	Fear	Hopeless	DDD	Lack of Info	Protecting	Age
Shame	0.33	0.13	0.18	0.19	0.11	0.09	0.12	0.09
Denial	0.06	0.06	0.03	0.11	0.03	0.08	0.08	0.05
Fear	0.20	0.12	0.28	0.23	0.07	0.13	0.10	0.19
Hopelessness	0.24	0.18	0.22	0.58	0.14	0.09	0.15	0.13
DDD	0.06	0.02	0.02	0.05	0.02	0.04	0.02	0.03
Lack of Info	0.06	0.06	0.06	0.09	0.07	0.19	0.02	0.13
Protecting	0.01	0.01	0.04	0.03	0.01	0.02	0.07	0.01
Age	0.07	0.03	0.08	0.08	0.01	0.12	0.03	0.11

2. Balanced

To combat the fact that the algorithms were not producing good results across all classes, the data was split into 3 groups to allow for more balanced data. The three groups were determined by how much data was present for each category.

(a) Reasons with Low Representation

The first group includes drugged, drunk, or disassociated and protecting assailant because both did not have many tweets associated with them. Thus they presented lower amounts of testing and training data. When they were separated from the whole group, the results were much better. The precision raised to 88% at max for Naive Bayes.

Table 13: Naive Bayes, Balanced, Multi-Class, Feature: Count

	precision	recall	f-score
Drugged, Drunk, Disassociated	0.85	0.91	0.88
Protecting Assailant	0.84	0.74	0.79
micro avg	0.84	0.84	0.84
macro avg	0.84	0.83	0.83
weighted avg	0.84	0.84	0.84

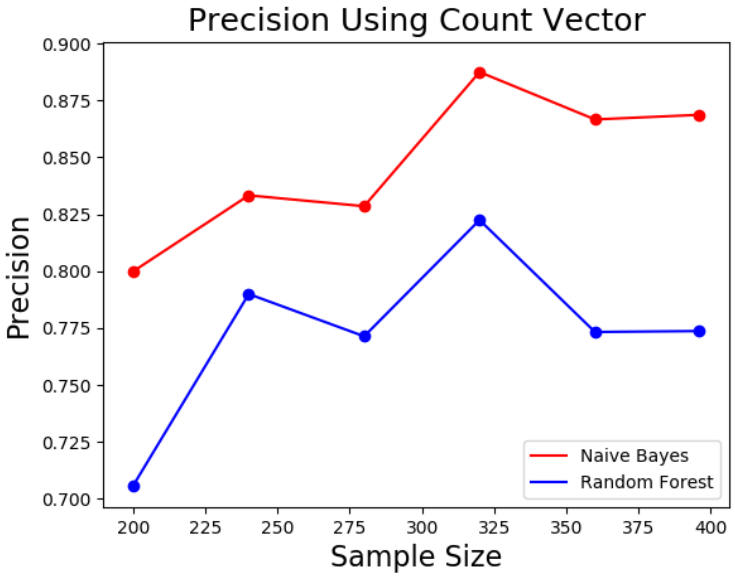

Figure 16: The graph portrays the precision of the multi-class classification algorithms using count vectors on the reasons mentioned the least.

(b) Reasons with Medium Representation

The second group includes denial/minimization, lack of information and age because each had enough data but not an overwhelming amount. When separated from the whole group, the results were better as expected but did not present a shocking difference. Most likely due to the fact that even in the big group these classes were generally labeled correctly.

Table 14: Naive Bayes, Balanced, Multi-Class, Feature: Count

	precision	recall	f-score
Denial/Minimization	0.68	0.64	0.66
Lack of Information	0.55	0.64	0.59
Age	0.63	0.55	0.59
micro avg	0.61	0.61	0.61
macro avg	0.62	0.61	0.61
weighted avg	0.61	0.61	0.61

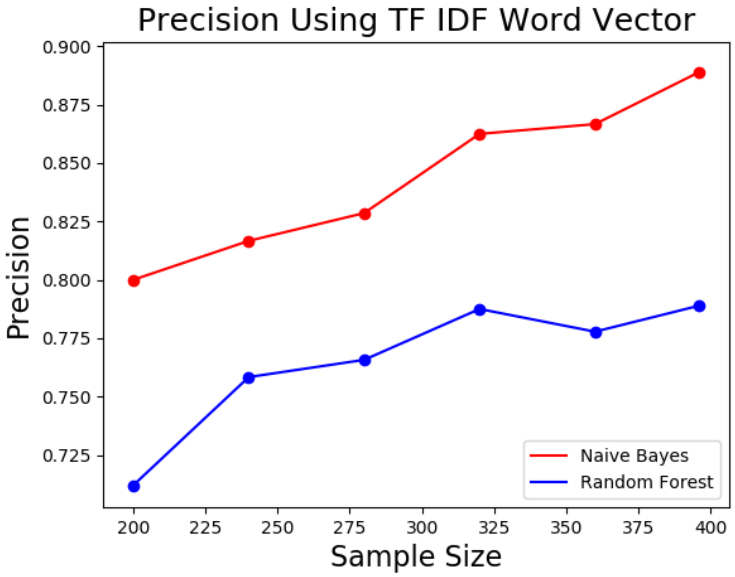

Figure 17: The graph portrays the precision of the multi-class classification algorithms using TF-IDF word vectors on the reasons mentioned the least.

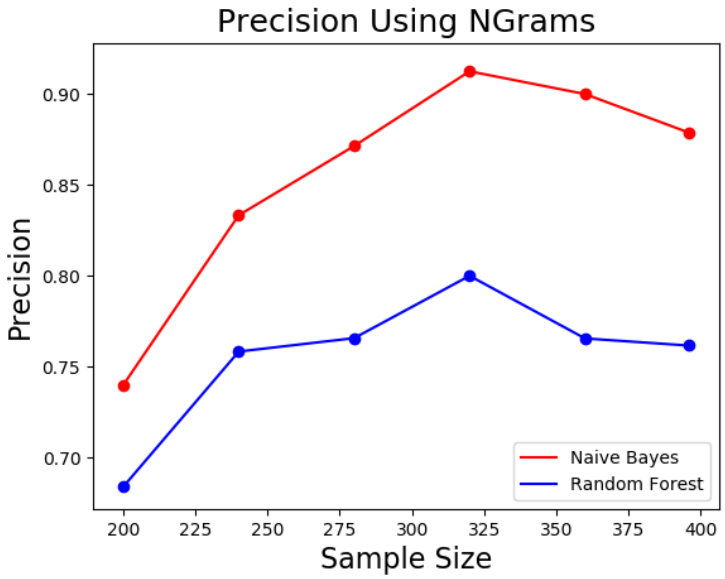

Figure 18: The graph portrays the precision of the multi-class classification algorithms using ngrams on the reasons mentioned the least.

(c) Reasons with High Representation

The third and final group includes shame, fear of consequences and hopelessness/helplessness because these three combined represent the majority of tweets. When separated from the whole group, the results were again better, but like the second group did not present a

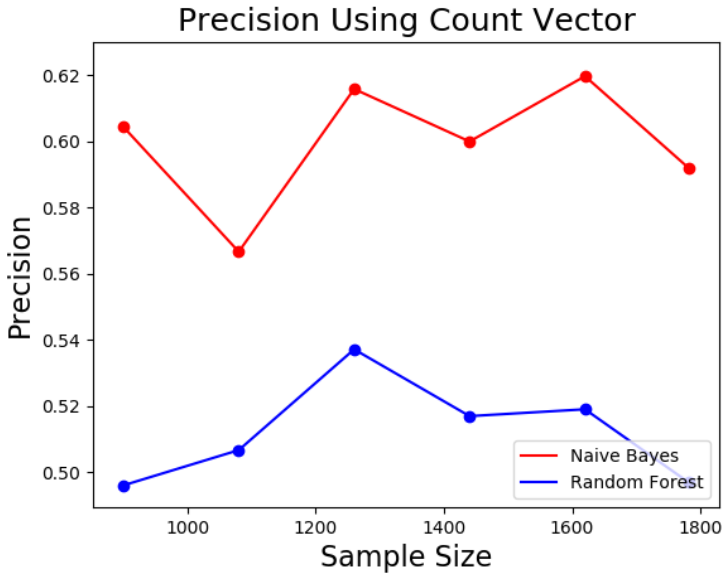

Figure 19: The graph portrays the precision of the multi-class classification algorithms using count vectors on the reasons mentioned somewhat often.

shocking difference. These classes affected the average the most when in the large group, so it is not surprising to see a smaller jump in performance.

Table 15: Naive Bayes, Balanced, Multi-Class, Feature: Count

	precision	recall	f-score
Shame	0.54	0.54	0.54
Fear	0.53	0.47	0.50
Hopelessness	0.65	0.71	0.68
micro avg	0.59	0.59	0.59
macro avg	0.58	0.57	0.57
weighted avg	0.59	0.59	0.59

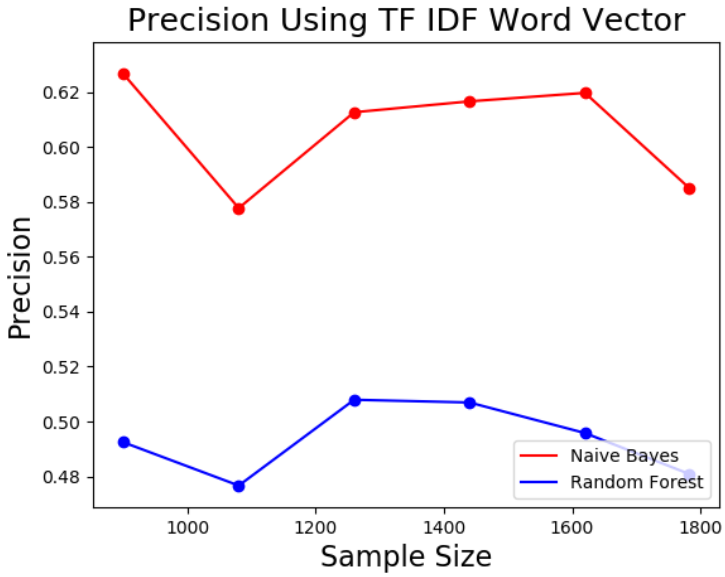

Figure 20: The graph portrays the precision of the multi-class classification algorithms using TF-IDF word vectors on the reasons mentioned somewhat often.

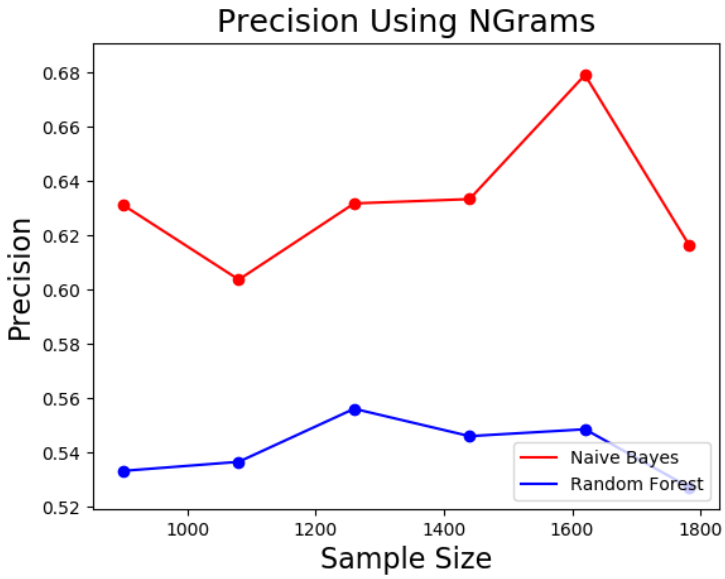

Figure 21: The graph portrays the precision of the multi-class classification algorithms using ngrams on the reasons mentioned somewhat often.

Splitting the data provided better results for all categories. The ability to focus on classes and putting classes with similar data set sizes together provided for better overall performance of all of the algorithms.

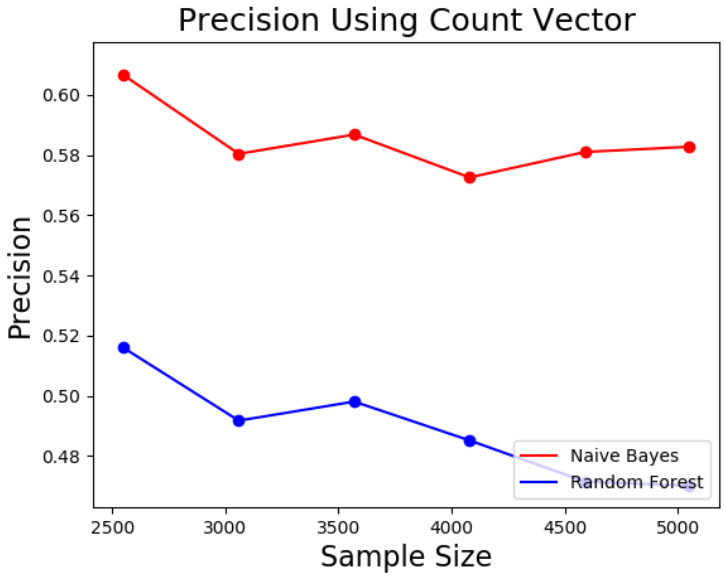

Figure 22: The graph portrays the precision of the multi-class classification algorithms using count vectors on the reasons mentioned the most.

4.2 Distribution of Reasons

4.2.1 Manually Labeled

The distribution of the reasons was different than that of other literature. A study using a survey of college women, [Spencer et al.2017], found that the most common reasons were lack of information and denial or minimization. The difference could stem from self-reporting vs a survey or from looking at a specific subset. Table 16 details the distribution they found based on a survey of college women [Spencer et al.2017]. The actual titles of the reasons have been adjusted to match the ones I used, but the percentages are all from [Spencer et al.2017]. The college-level study did not take into account age, as all of the women were about the same age when the assault occurred.

Table 16: [Spencer et al.2017] Distribution of Reasons

Reason	Paper's Percent	My Percent
Shame	8	26.49
Denial/Minimization	28.4	9.85
Fear of Consequences	10.2	28.88
Hopelessness/Helplessness	4.8	31.14
Drugged, Drunk, or Disassociated	8.8	5.8
Lack of information	18.6	12.03
Protecting Assailant	7.1	3.6
Age	0	10.73

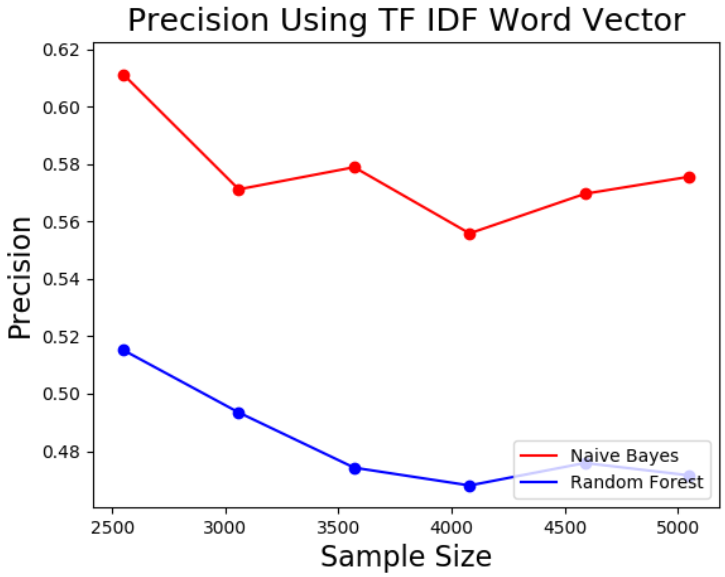

Figure 23: The graph portrays the precision of the multi-class classification algorithms using TF-IDF word vectors on the reasons mentioned the most.

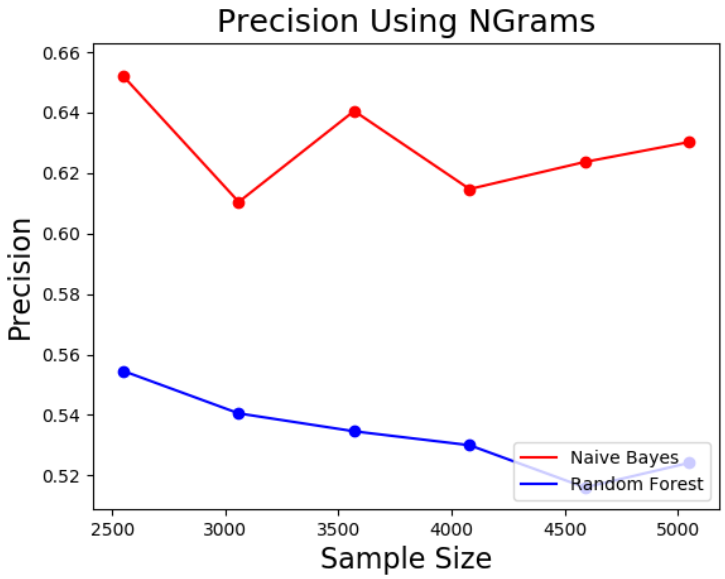

Figure 24: The graph portrays the precision of the multi-class classification algorithms using ngrams on the reasons mentioned the most.

A national survey of college women also had different results [Fisher et al.2003]. The study was a little harder to directly compare, as the reasons studied were more focused than the ones picked for this thesis. The National College survey mainly looked at Fear of Consequences, Denial/Minimization, and Hopelessness/Helplessness. Some other categories were presented, but they

Table 17: [Fisher et al.2003] Distribution of Reasons

Reason from Paper	The Reason Corresponding to this Thesis	Paper's Percent
Not Want family to Know	Fear	18.3
Not Want others to Know	Fear	20.9
Not Sure it was Intentional		42.1
Lack of Proof	Hopelessness	23.2
Not Serious enough	Denial	81.7
Police Not Think Serious Enough	Hopelessness	28.6
Not Want to Bother Police		21.3
Fear of Treatment by Cops	Fear	7.8
Fear of Treatment in Judicial Process	Fear	3.0
Fear of Reprisal	Fear	19.0
Not Know How to	Lack of Information	8.1
Other		2.3

did not correspond to any in this study. Again, the results revealed a much higher presence of Denial/Minimization than was discovered in this thesis. Table 17 details how the findings of [Fisher et al.2003] and this thesis correspond. The main thing to notice is that the reasons were far more specific in [Fisher et al.2003] which aligns with the overall purpose of both that study and this thesis. This thesis was seeking to make a more general conclusion by using social sensing and Twitter. While the survey had the goal of examining a specific group and more specified reasons.

The Criminal Justice System gathered statistics at a national level that are presented in [Langton and Truman2015 It had different reasons that were not as specific but has a much closer spread of why people did not report to what was discovered in this thesis. Fear and belief of nothing being done were two of the most common reasons, just as was the case from the Twitter data. The close relation could be a result of the data being gathered on a large scale. The survey still does not detail the same reasons and leaves out shame which was a common and important reason discovered using Twitter.

Table 18: [Langton and Truman2015] Distribution of Reasons

Reason from Survey	Reason Corresponding to Thesis	Paper's Percent	Thesis Percent
Fear of Retaliation	Fear	20.0	28
Police Won't help	Hopelessness	15	31
Personal		13	
Did Report		8	
Not Important Enough	Denial	8	9.8
Protect Perpetrator	Protecting Assailant	7	3.6
Other/No Reason		30	

4.2.2 Machine Learning Algorithm Labeled

In order to label the data using the machine learning algorithms, I used both the binary and multi-class algorithms. I first pulled the 30,000 tweets that were not originally manually labeled. These tweets were labeled as having a reason or not by using the RNN classification algorithm. The tweets that contained a reason were passed to the large data multi-class algorithm. The large data algorithm was chosen because these three classes represent 70% of the tweets, and this algorithm was more precise than the algorithm classifying using all reasons. Figure 25 portrays the distribution of reasons produced using this chain of events.

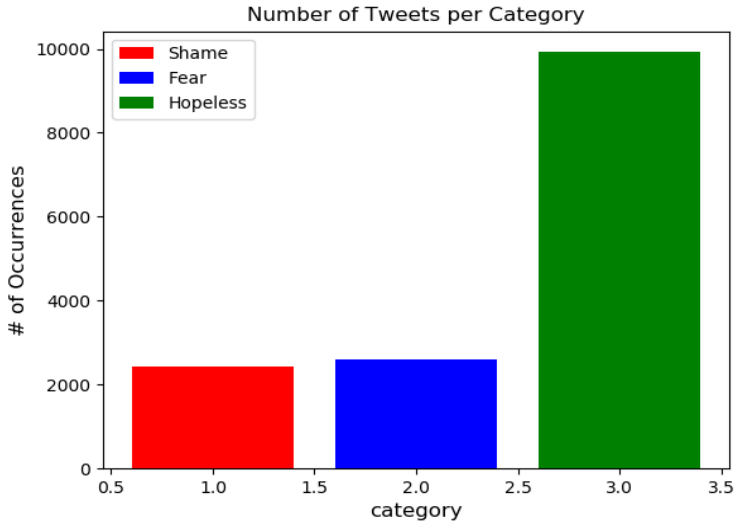

Figure 25: The bar chart displays the distribution of reasons generated from the algorithms classifying the tweets.

5 Discussion

5.1 Reasons for Not Reporting

The opportunity to use data from Twitter presents a unique study where all data is self-reported. It resulted in a difference in the reasons that were most prominent in victims not reporting their assault as compared to other forms of data collection. Self-reporting prevents clarification opportunities. Interviews and surveys can allow for more pinpointing of what the reason was, while self-reporting allows for no real background on the scenario.

Victims are able to voice their thoughts and frustrations on a platform and in a movement where they feel heard. I believe this led to a higher appearance of victims sharing that they did not report because of a belief, fear or evidence that nothing will be done if they did come forward. Twitter provides a voice to people who feel voiceless, while a survey or interview provides no publicity.

The movement also lends itself to people being more likely to express frustration since it was born from victims not being believed. Since #WhyIDidntReport began to support Dr. Christine Blasey Ford, a victim being accused of lying, it is more likely that other victims will cite the same issue as it most closely aligns with the events that sparked the movement. This is partially a result of Twitter being an echo chamber, which is usually applied to politics but can be generalized to apply to any topic on Twitter. Users are more likely to see and interact with ideas they agree with, in doing so, it creates this "echo" of more of the same things they agree with [Garimella et al.2018]. In addition to this, Twitter provides an opportunity for people to share stories they are not always comfortable sharing in real life but only allows brief descriptions. Because of this, it can result in more engagement with other users causing similar reasons to repeat [Manikonda et al.2018].

5.2 Twitter

Social movements allow for many to voice their opinions. The findings of this study were not consistent with those of studies that collected data using other methods. Self-reporting on a public platform results in users being influenced by what others think or say. After reading the tweets, many users applied similar wording to their references, thus possibly implying being influenced by other users and other tweets they have read. Unlike an interview where questions can lead to specific results, self-reporting leaves the response open-ended, but to believe that other people's tweets have no effect on other users would be naive. The results could also be influenced by what triggered the movement.

The Kavanaugh hearings exemplify a scenario in which the victim's testimony was not met with positive reactions by all. Many tweets reference not reporting because of a belief that they will not be believed or nothing will be done. The hearings may have only solidified this fear. Because of this, it makes sense that many tweets express frustration with the justice system and society for how they react to accusations. Unlike interviews and surveys, social media is heavily affected by society, reactions, and perceptions. Twitter offers a great avenue to see just how much people's perceptions of themselves and their circumstances are affected by others on a large scale.

5.3 Contribution to Future Research

The main portion of this study to be used in the future are the algorithms and labeled data set. Using the labeled data and machine learning algorithms, other researchers can label large sets of data automatically rather than manually. The vast amount of data presents the opportunity to explore the social culture surrounding this topic on an even greater scale. Studies can be conducted on the language surrounding specific topics to pinpoint exactly how to make victims feel safe in coming forward.

The opportunity to continue to study and dissect the reasons victims do not report is important in understanding how to change the culture surrounding reporting sexual assault. It can also be a good resource for psychology research as a reference for why people do not report as discovered on a large scale. The use of Twitter, machine learning algorithms, and a popular topic plaguing society today in this thesis presents a great starting place and reference for various forms of research.

6 Limitations

While Twitter provides a huge amount of data, in order to generate the training and testing data, I had to manually label the data. The amount of time labeling the data limits the ability to reach a sufficiently large sample size for each category. As a result, RNN was not able to be used to label which reason was cited in the tweets, since it requires much more data to produce accurate results. It also limited the ability to see how accurate Naive Bayes and Random Forest could be with larger amounts of data since more data could not be generated in a reasonable amount of time.

Twitter's large amount of data was a great way to explore new ways of answering this question, but the effort behind labeling the data presented a limitation. Because the subject can be rather emotional and subjective, labeling tweets with which reason the victim cited allows for bias. To combat that, I would have liked to have a few more people label the data in order to make sure the labels were accurate. However, because labeling took such a large amount of time, a smaller, consistently labeled set was deemed preferable.

The categorizing also assumes that all accounts are in fact sexual assault or harassment. In any survey, interview, or empirical form of gathering data, the researcher can clarify exactly what they mean by sexual assault. Because the data is self-reported, there is no pinpointing of exactly what constitutes sexual assault, but this does allow for victims to determine for themselves what felt inappropriate to them. Also, without clarifying this, it could be the reason for fewer appearances of denial or minimization, since explaining what falls into the category of assault may bring to someone's attention that they were assaulted and just chose to view it as not a big enough deal.

Finally, Twitter only allows 280 characters per tweet. The fact that tweets are so limited, can prevent a user from going into too much detail. Some may argue that the limitation requires more thought put into each word, but the reality is it probably just results in some details being left out. The physical limit on how much someone could write about their experience presents limited

information thus affecting how much understanding can be had about each circumstance.

7 Expanding on this Study

To build on this study, one could generate a larger sample. Taking the time to label more data would allow for other machine learning algorithms to be used and theoretically produce more accurate results. It also could reduce the bias that comes into play from having a person make the most informed decision about what category the victim is referencing in their tweet as to why they did not report.

Aside from collecting more data, it would be ideal to have more subjects label the tweets. Each subject should label if a reason is mentioned and what the reason is that both can have a level of personal bias and opinion. It could also be important to have them label large sets of data because as you label more data, you begin to get a better sense of what tweets go with which category, at least in your opinion.

It would also present an interesting opportunity to compare this study with another self-reporting study. This could be done through comparison of results from other social media platforms, as these are all truly self-reporting. The #WhyIDidntReport movement did call victims to action, but aside from bringing awareness to the issue, the data present in this study were all victim supplied with no one directed where the response should go. Because of this, it would be important to consider data that was also supplied without much direction from a researcher.

8 Conclusion

Twitter allowed for an interesting study that revealed differences in the reasons victims do not report assault when data is gathered through self-reporting rather than a survey or interview. The data revealed a higher tendency to say they did not report because they did not feel it would help. While surveys revealed a higher percentage of victims claiming denial or minimization as their reason. Both reasons were considered in this study, but hopelessness or helplessness was the leading explanation.

The machine learning algorithms were used to get a basic understanding of the data and to test out the precision of a few on this type of data. Recurrent Neural Networks outperformed Naive Bayes and Random Forest on binary classification but was not used for multi-class classification. Naive Bayes was the most precise at determining the reasons mentioned in the tweets.

BIBLIOGRAPHY

[Carney2016] Carney, N. (2016). All lives matter, but so does race: Black lives matter and the evolving role of social media. *Humanity & Society*, 40(2):180–199.

[Engel2017] Engel, B. (2017). Why don't victims of sexual harassment come forward sooner? *Psychology Today*.

[Fisher et al.2003] Fisher, B. S., Daigle, L. E., Cullen, F. T., and Turner, M. G. (2003). Reporting sexual victimization to the police and others: Results from a national-level study of college women. *Criminal justice and behavior*, 30(1):6–38.

[Garimella et al.2018] Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2018). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 913–922. International World Wide Web Conferences Steering Committee.

[Ince et al.2017] Ince, J., Rojas, F., and Davis, C. A. (2017). The social media response to black lives matter: how twitter users interact with black lives matter through hashtag use. *Ethnic and racial studies*, 40(11):1814–1830.

[Khan et al.2018] Khan, S. R., Hirsch, J. S., Wambold, A., and Mellins, C. A. (2018). 'i didn't want to be'that girl": The social risks of labeling, telling, and reporting sexual assault. *Sociological Science*, 5:432–460.

[Langton and Truman2015] Langton, L. and Truman, J. (2015). Criminal victimization, 2014. *Bureau of Justice Statistics, US Department of Justice*.

[Manikonda et al.2018] Manikonda, L., Beigi, G., Liu, H., and Kambhampati, S. (2018). Twitter for sparking a movement, reddit for sharing the moment: #metoo through the lens of social media. *arXiv preprint arXiv:1803.08022*.

[Ménard2005] Ménard, K. S. (2005). *Reporting sexual assault: A social ecology perspective*. LFB Scholarly Pub.

[Mengeling et al.2014] Mengeling, M. A., Booth, B. M., Torner, J. C., and Sadler, A. G. (2014). Reporting sexual assault in the military: who reports and why most servicewomen don't. *American journal of preventive medicine*, 47(1):17–25.

[Noveck2018] Noveck, J. (2018). Shame, fear, power: Hitting back at trump's tweet, survivors explain why they didn't report sexual assaults. *Business Insider*.

[O'Neil et al.2018] O'Neil, A., Sojo, V., Fileborn, B., Scovelle, A. J., and Milner, A. (2018). The #metoo movement: an opportunity in public health? *The Lancet*, 391(10140):2587–2589.

[Rohrer2017] Rohrer, B. (2017). Recurrent neural networks (rnn) and long short-term memory (lstm).

[Soni2018] Soni, D. (2018). Introduction to naive bayes classification.

[Spencer et al.2017] Spencer, C., Mallory, A., Toews, M., Stith, S., and Wood, L. (2017). Why sexual assault survivors do not report to universities: A feminist analysis. *Family relations*, 66(1):166–179.

[Ullman2016] Ullman, S. E. (2016). Sexual revictimization, ptsd, and problem drinking in sexual assault survivors. *Addictive behaviors*, 53:7–10.

[Ullman and Peter-Hagene2014] Ullman, S. E. and Peter-Hagene, L. (2014). Social reactions to sexual assault disclosure, coping, perceived control, and ptsd symptoms in sexual assault victims. *Journal of community psychology*, 42(4):495–508.

[Ward2018] Ward, S. F. (2018). Times up: As the me too movement continues to shed light on sexual harassment and assault, sparking changes in various industries, the legal and judicial systems have been slow to adapt. *ABA*, 104(6):46–54.